Can we Trust Explanation! Evaluation of Model-agnostic explanation techniques

Syed Ihtesham Hussain Shah[1][0000−0002−6390−1864], Annette ten Teije[1][0000−0002−9771−8822], and Jose Volders[2]

[1] Faculty of Science, Department of computer science, Vrije Universiteit Amsterdam, Netherlands {s.i.h.shah,annette.ten.teije}@vu.nl
[2] Diakonessenhuis, Netherlands
voldersjh@gmail.com

Abstract. Explainable AI (XAI) assists clinicians and researchers in understanding the rationale behind the predictions made by data-driven models which helps them to make informed decisions and trust the model's outputs. However, given the variety of explanation techniques, there is no universally applicable evaluation metric that can reliably assess the quality of all explanations. This study addresses this gap by introducing a set of universal evaluation metrics designed to assess explanation performance across different techniques and contexts. We conduct a comprehensive comparison of two widely used post-hoc explanation methods: Local Interpretable Model-Agnostic Explanations (LIME) and SHapley Additive Explanations (SHAP) applied to a highly imbalanced multiclass-multioutput breast cancer treatment prediction task. These methods were evaluated using proposed evaluation matrices that included fidelity, stability, consistency, and alignment with clinical guidelines. Our findings reveal that SHAP generally provides more faithful and consistent explanations than LIME, especially in alignment with clinical knowledge. These results reinforce the need for tailored evaluation strategies rather than relying on a single universal metric, highlighting that the choice of explanation method should be informed by the specific clinical context and interpretability goals.

Keywords: Explainable AI · Black-Box · LIME · SHAP · Breast Cancer

1 Introduction

In the healthcare industry, XAI is used frequently to manage clinical diagnosis [8], drug delivery [2], disease classification and treatment recommendations [4] [7]. Some studies [3, 6] highlight the strengths and weaknesses of two widely used post-hoc explanatory methods Local Interpretable Model-agnostic Explanation (LIME) [1] and SHapley Additive exPlanation (SHAP) [5]. However, to the best of our knowledge, direct comparisons of LIME and SHAP using the same evaluation metrics, especially in healthcare domains remain limited. The primary objective of this research is to propose a set of universally applicable evaluation metrics and to conduct an in-depth comparison of LIME and SHAP

for predicting breast cancer treatments. For a fair comparison of the explanation techniques, we introduced both Application-level and Human-level evaluations. Application-level evaluation consists of assessing fidelity, stability, and alignment with medical guidelines, while Human-level evaluation involves a qualitative analysis of the generated explanations, focusing on their interpretability and usefulness from an expert's perspective.

Fidelity measures the similarity of the prediction made by a black box and a surrogate model, it can be represented as: $R^2 = 1 - \frac{\sum_{i=1}^{k}(f(z^{(i)}) - g(z^{(i)}))^2}{\sum_{i=1}^{k}(f(z^{(i)}) - \bar{f})^2}$. Where $f(z^{(i)})$ are predictions for perturbed samples from the complex model, $g(z^{(i)})$ are predictions for perturbed samples from the surrogate model and $\bar{f}$ is the mean of the original model's predictions.

Stability compares the variables composition in the explanations that are generated multiple times for the same instance. $Stability = \frac{\sum_{1}^{k} \frac{C_{pair}}{p}}{|C_m^2(\mathcal{E}^1,\ldots,\mathcal{E}^m)|}$. Where concordance function $\mathcal{C}_{pair}$ returns cardinality of the intersection between two explanations, p is the number of variables in the explanations and C_m^2 is the pair in the explanation $\mathcal{E}^i$.

Comparison with guidelines: Let G is the set of medical guidelines, where $G_i \in \{High,\ Medium, Low\}$ represents the importance of feature i. Comparison index $\Gamma(G, \mathcal{E}), measurestheconcordancescoresbetweenexplanations\mathcal{E}$ and medical guidelines G, can be defined as: $\Gamma(\mathcal{E}, G) = \frac{1}{|G|}\sum_{i=1}^{|G|}\mathcal{I}_i$. A high value of $\Gamma(\mathcal{E}, G)$ close to 1 indicates that the explanations are completely matches the guidelines and vice versa.

Human-Level Evaluation: We conducted survey with the clinician who evaluated the explanations based on seven key criteria: Understandability, Satisfaction, Level of Detail, Completeness, Trustworthiness, Predictability, and Safety/Reliability. Each criterion was rated on an integer scale from 1 (very poor) to 10 (excellent), providing a quantitative measure of the perceived quality of the explanations.

2 Conclusion

In this paper, we introduced a set of metrics to assess the quality of any model-agnostic explanation technique, regardless of its underlying working principle. We focused on two widely used explanation methods, LIME and SHAP, which operate on fundamentally different principles, and conducted a comprehensive analysis of their performance in predicting breast cancer treatments using a highly imbalanced synthetic IKNL dataset. Our experiments showed that SHAP outperformed LIME in terms of fidelity, stability and and more consistently aligned with medical guidelines and with the expert evaluation than LIME. For additional information, please see the full article at the following link.
https://www.scitepress.org/Papers/2025/131574/131574.pdf

Bibliography

[1] Touhidul Islam Chayan, Anita Islam, Eftykhar Rahman, Md Tanzim Reza, Tasnim Sakib Apon, and MD Golam Rabiul Alam. Explainable ai based glaucoma detection using transfer learning and lime. In *2022 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)*, pages 1–6. IEEE, 2022.

[2] José Jiménez-Luna, Francesca Grisoni, and Gisbert Schneider. Drug discovery with explainable artificial intelligence. *Nature Machine Intelligence*, 2 (10):573–584, 2020.

[3] I. Elizabeth Kumar, Suresh Venkatasubramanian, Carlos Scheidegger, and Sorelle Friedler. Problems with shapley-value-based explanations as feature importance measures, 2020.

[4] Monika S Mellem, Matt Kollada, Jane Tiller, and Thomas Lauritzen. Explainable ai enables clinical trial patient selection to retrospectively improve treatment effects in schizophrenia. *BMC medical informatics and decision making*, 21(1):162, 2021.

[5] Yuan Meng, Nianhua Yang, Zhilin Qian, and Gaoyu Zhang. What makes an online review more helpful: an interpretation framework using xgboost and shap values. *Journal of Theoretical and Applied Electronic Commerce Research*, 16(3):466–490, 2020.

[6] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier, 2016.

[7] Syed Ihtesham Hussain Shah, Giuseppe De Pietro, Giovanni Paragliola, and Antonio Coronato. Projection based inverse reinforcement learning for the analysis of dynamic treatment regimes. *Applied Intelligence*, 53(11):14072–14084, 2023.

[8] Yiming Zhang, Ying Weng, and Jonathan Lund. Applications of explainable artificial intelligence in diagnosis and surgery. *Diagnostics*, 12(2):237, 2022.